Towards explained treatment search results: feature analysis and explanation formulation

Contempré, Edeline; Szlávik, Zoltán; Velazquez Godinez, Erick; ten Teije, Annette; Tiddi, Ilaria

2021

document version
Peer reviewed version

Towards explained treatment search results: feature analysis and explanation formulation

Edeline Contempré[1,2][0000−0002−1767−121X], Zoltán Szlávik[1][0000−0002−2781−3795], Erick Velazquez Godinez[1][0000−0001−9449−8265], Annette ten Teije[2][0000−0002−9771−8822], and Ilaria Tiddi[2][0000−0001−7116−9338]

[1] myTomorrows, Anthony Fokkerweg 61, 1059 CP Amsterdam, The Netherlands {edeline.contempre,zoltan.szlavik,erick.velazquez}@mytomorrows.com
[2] Vrije Universiteit Amsterdam, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands {annette.ten.teije,i.tiddi}@vu.nl

Abstract. We are presenting preliminary results of a work-in-progress project that aims to increase healthcare professionals' trust in treatment search engines by introducing a) explainability based re-ordering of retrieved documents, and b) providing user-friendly explanations for each of these documents. Through the use of crowdsourcing, we assess the importance of various features for explainability, and also investigate aspects of explanation formulation as presented to end-users. Our results allow us to determine feature weights that will be inputs to the document re-ordering model, the per-document explanatory sentence formulation module, and the sentence ordering model.

Keywords: Explainability · Feature analysis · Healthcare · Treatment search.

1 Introduction

Healthcare professionals (HCPs) increasingly rely on Artificial Intelligence (AI) models to help their patients and save their lives. This could especially be the case for HCPs referring patients (with often short prognosis) to clinical trials as, for these patients, no treatments might be available on the market. While various relevant AI models' accuracy keeps increasing, their underlying processes, and consequently, their outcomes, are becoming increasingly difficult to understand. In the medical domain, where the pressure to make no mistakes is high, not being sure why and how a decision is made could have dire consequences for everyone involved. Consequently, without understanding an AI model and its output, HCPs' trust in the model will decrease, potentially leading to abandoning its use [1].

In this paper, we present the first results from a work-in-progress project in which we aim to develop a local explainability [2] based method, that a) results in a model that, based on explainability, **re-orders documents** retrieved by a treatment search engine, and b) provides an ordered list of **explanatory sentences** to end-users (HCPs) for each retrieved document. We believe that,

through enabling HCPs to efficiently find relevant **clinical trials** while understanding search results, we can increase their trust in the search engine, and ultimately facilitate the process of getting more patients treated.

The paper focuses on the first stage of the research, which – through the use of crowdsourcing – identifies and assesses i) the importance of features used for explainability, and ii) the formulation of explanations as presented to end-users.

2 Explainability features, in context

Our method will generate local explainability scores for each result retrieved by a search engine, and use these scores to re-order the search engine's results. For each clinical trial in the result list, users will also be shown user-friendly formulated sentences as explanations providing descriptions of features playing a role in the retrieval process. We rely on a set of user-interpretable features for both re-ordering and explanatory sentences (see Figure 1 for the pipeline).

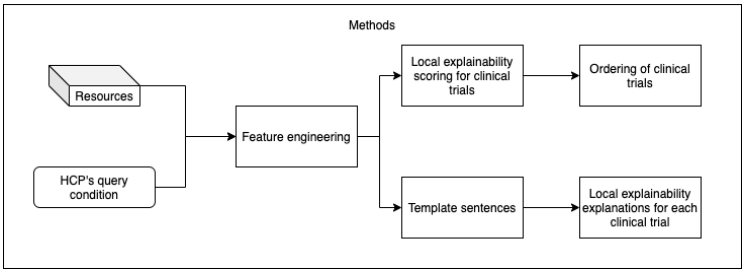

Fig. 1. Explainability pipeline overview.

We determined the individual features based on related work on features used for learning-to-rank (e.g., [3]), the LIRME method [4], earlier work within the company [5], and conversations with company UX designers and stakeholders. Features can be classified into categories based on whether their value depends on the search query, and their output type (i.e. binary – receives a score of 1 if present or 0 if not, or numeric – a count that will be normalised). An overview of all selected features is shown in Table 1. For example, the feature *Query in title* is a query dependent feature as its value depends on a match between the query (e.g. breast cancer) and the title. In contrast, *Number of publications* associated with a clinical trial remains the same regardless of the user's query, and can take up a value larger than one.

	Query dependent	**Query independent**
Binary	Query in title, Preferred term in title	Clinical stage present, Stage is recruiting, Overall status given
Numeric	Query in summary, Preferred term in summary, Preferred term in summary, Query in detailed description	Number of publications

Table 1. Classification of features created for the local explainability based search engine.

Our explainability method does not depend on the exact method used for the search engine retrieving relevant trials, though naturally there is overlap between what both methods make use of. The search engine used for this research[3] extracts properties of clinical trials such as the trial's title, trial stage, summary, etc., and uses these to match trials to queried medical conditions. The extraction process relies on a condition-focused knowledge graph based on UMLS[4] and Disease Ontology[5]. We also use PubMed[6] abstracts for some features, and process about 400.000 trials from clinical trial registries such as clinicaltrials.gov[7].

3 Crowdsourced feature importance

In order to combine features to eventually re-order search results, we decided to determine features' weights based on their perceived importance, through their corresponding explanatory sentences, by users. Weights for features may also be used to rank explanatory sentences for each retrieved trial.

As no data was available on users' interaction with the search engine, and no real-time user-interaction data could be collected for privacy reasons, we were not able to infer weights through search engine usage. Our solution was to use an implicit *preference elicitation strategy* via crowdsourcing, as – although HCPs were the target audience for this research – earlier studies employing crowdsourcing tasks have shown that lay participants' annotation reach similarly high levels of quality output compared to that of niche-sourcing with medical experts [6, 7].

Our hypotheses were twofold: a) features are not equally preferred by users, and b) formulation of explanations makes a difference. More specifically, we are interested in three dimensions of sentence formulation: numeric vs. non-numeric (*'5 times'* vs. *'multiple times'* mentioned in explanatory sentences), action-oriented vs. fact-driven formulations (*'retrieved'* vs. *'clearly mentioned'*), and disease specific vs. non-disease specific outputs (*'HIV'* vs. *'condition'*).

To address these, we created a labeling job that simulated the context of the search engine of myTomorrows, and ran these via Amazon Sagemaker Ground Truth (which uses Amazon Mechanical Turk). Crowd workers were asked to imagine themselves in a situation where they had a friend or family member with one of the four randomly allocated conditions: *invasive breast cancer*, *Lyme disease*, *HIV*, or *chronic migraine*. Furthermore, they were asked to imagine themselves querying the condition in a search engine containing clinical trials, and to rate sentences based on how convincing a sentence was to make them want to read clinical trial in further detail and more closely assess its relevance. Note that due to the difficulty of understanding clinical trial documents by lay people [8], participants were not asked to look at such documents.

[3] https://search.mytomorrows.com/public
[4] https://www.nlm.nih.gov/research/umls
[5] https://disease-ontology.org/
[6] https://pubmed.ncbi.nlm.nih.gov/
[7] https://www.clinicaltrials.gov/

We created different sentences for the four conditions, sentences for the 10 features (see table 1), using the three dimensions of formulation, and presented these to workers in a random order. In addition, we employed 'gold questions' used in crowdsourcing to control for quality.

4 Results and discussion

Through crowdsourcing, we obtained 1116 responses to our labeling task. Each feature was attributed 16 sentences where each single sentence's task was answered by 9 workers at a time (the maximum number per task on the platform). This way we collected n = 144 data points per feature. *Numeric*, *non-numeric*, *retrieve* and *clearly mentioned* entities were each assessed by n = 144 workers. *Disease specific* and *disease unspecific* entities were each assessed by n = 198 workers. 10% of workers answering the gold question were discarded as suspected low quality workers.

Data obtained with respect to **feature importance** shows that there is at least a partial ordering that can be obtained via the crowdsourcing (based on statistical tests). The feature with the highest mean score (3.69, on a 5-point Likert scale) was *Query in detailed description*, whereas the two least convincing features were *Query in title*, and *Trial is recruiting* (3.15, and 3.13, respectively). Our interpretation of these results is that features that people accept as more intuitive bring less convincing power, i.e. explainability: through web search, people are used to query words appearing in the title, and people tend to be looking for trials that are recruiting at the moment. We determined the features' weights, using chi-square tests, based on statistical values. If two features were, for example, not statistically equally preferred, these two features would be attributed different weights.

When it comes to results for the three **formulation** dimensions (Table 2), when performing chi-square tests, we found that: 1. users prefer explanations without specific numbers mentioned (e.g. query term occurring *'multiple times'* vs. *'5 times'*). This may indicate that users are used to words and are, therefore, more convincing than those that require more mental processing, i.e. when specific numbers are displayed. 2. Users seem to prefer facts (*'clearly mentioned'*) to actions related to the search procedure (*'retrieved'*). 3. There is no preference difference between being specific about the condition in the query (e.g. *'HIV'*) vs. (*'the condition'*). This is in line with the finding in the previous paragraph that what is obvious, i.e. that we are looking for documents containing the query term, need not be mentioned in detail in an explanation.

The results provided us with guidance for a) which features are considered more important in re-ordering search results based on explainability, b) the formulation of explanatory sentences, and c) ordering explanatory sentences for individual retrieved documents. The formulation point is particularly interesting for future work, as it suggests that a considerable amount of research is still needed on how search experience should be designed. To the best of our knowledge, no related work has been done on the type of explanations users would like to see associated with retrieved documents in the medical domain.

Entity	$\overline{x}$(1)	$\overline{x}$(2)	P-value
(1) Non-numerical	3.7	3.34	***0.01****
(2) Numerical			
(1) Clearly mentioned	3.65	3.33	***0.036****
(2) Retrieved			
(1) Specify disease	3.4	3.48	0.44
(2) Not specify disease			

Table 2. Preferences per formulation dimension (5-point Likert scales).

5 Conclusions and future work

In this paper we have briefly presented a preliminary study towards re-ordering documents retrieved by a treatment search engine, and providing end-user-friendly explanations for each document retrieved. We used crowdsourcing to assess both feature importance, and user preferences in terms of the formulation of explanatory sentences.

The next step in our work is to implement an 'explanation engine' attached to a search system, and measure the impact of both the re-ordering, and the explanations themselves. If privacy policy allows, we would, in addition, like to use results from this paper to address the cold start problem for an explanation engine that learns through user interaction.

References

1. R. H. Kay, "The relation between locus of control and computer literacy," *Journal of Research on Computing in Education*, vol. 22, no. 4, pp. 464–474, 1990.
2. A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52 138–52 160, 2018.
3. T. Joachims, "Optimizing search engines using clickthrough data," in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2002, pp. 133–142.
4. M. Verma and D. Ganguly, "LIRME: locally interpretable ranking model explanation," in *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2019, pp. 1281–1284.
5. S. Gu, "Improving search relevance feedback through human centered design," Master's thesis, TU Delft, The Netherlands, 2020.
6. A. Dumitrache, L. Aroyo, C. Welty, R.-J. Sips, and A. Levas, "Dr. Detective: combining gamification techniques and crowdsourcing to create a gold standard in medical text," in *Proceedings of the 1st International Conference on Crowdsourcing the Semantic Web*, vol. 1030, 2013.
7. A. Dumitrache, L. Aroyo, and C. Welty, "Crowdsourcing ground truth for medical relation extraction," *ACM Trans. Interact. Intell. Syst.*, vol. 8, no. 2, Jul. 2018. [Online]. Available: https://doi.org/10.1145/3152889
8. D. T. Wu, D. A. Hanauer, Q. Mei, P. M. Clark, L. C. An, J. Proulx, Q. T. Zeng, V. V. Vydiswaran, K. Collins-Thompson, and K. Zheng, "Assessing the readability of clinicaltrials. gov," *Journal of the American Medical Informatics Association*, vol. 23, no. 2, pp. 269–275, 2016.